Exercise set 1

Joseba Hernandez Bravo and Jorge Vicente Puig
17/12/2021

```
library(combinat)
```

TASK 1

In this first task we will examine the correlation between the pairs of (X1,Y1) values by means of a correlation test.

We will perform an exact test for $H_0 : \rho = 0$ against $H_1 : \rho > 0$ where:

- $H_0 : \rho = 0$: There is no correlation between Chest circumference and volume of air.
- $H_1 : \rho > 0$: Chest circumference and volume of air are positively correlated.

Correlation test

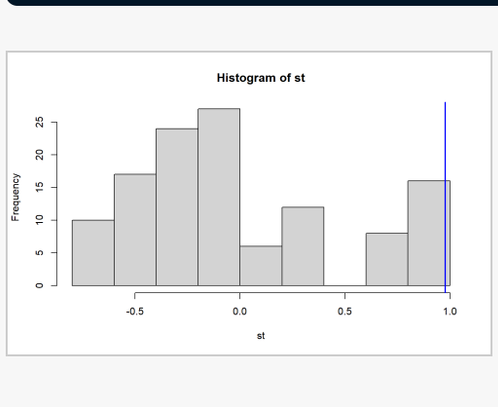
We will perform a permutation correlation test and plot an histogram and the p-value to analyse the results.

```
x= c(39,29,60,40,32)
y=c(11,5,20,8,6)

sttrue= cor(x,y)

n=length(y)
nr=fact(n) #number of rearrangements to be examined
st=numeric(nr)

cnt=0
d=permn(y)

for (i in 1:nr){
  st[i]<-cor(d[[i]],x)
  if (st[i] >=sttrue)
    cnt=cnt+1
}
```

```
print(paste("p-value= ",cnt/nr))
```

```
## [1] "p-value=  0.025"
```

```
hist(st)
abline(v=sttrue,col="blue",lwd=2)
```



As the p-value is less than 0.05 we will reject the null hypothesis. Then, we can say that there is sufficient evidence to conclude that the X and Y are correlated.

Pearson

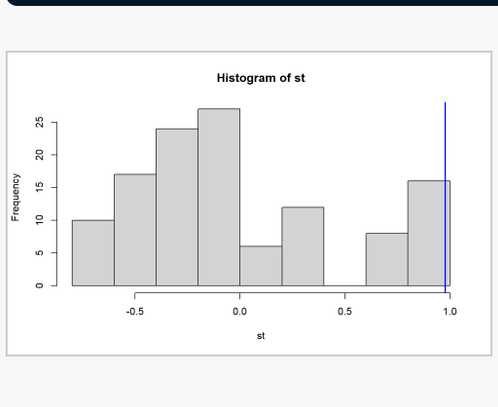
Now, we will perform the test using a Pearson correlation coefficient in order to have a different statistic for evaluating the null hypothesis.

```
x= c(39,29,60,40,32)
y=c(11,5,20,8,6)

sttrue= cor(x,y, method = "pearson")

n=length(y)
nr=fact(n) #number of rearrangements to be examined
st=numeric(nr)

cnt=0
d=permn(y)

for (i in 1:nr){
  st[i]<-cor(d[[i]],x)
  if (st[i] >=sttrue)
    cnt=cnt+1
}
```

```
print(paste("p-value= ",cnt/nr))
```

```
## [1] "p-value=  0.025"
```

```
hist(st)
abline(v=sttrue,col="blue",lwd=2)
```



As in the previous case, we have that our p-value < 0.05 so, we will reject the null hypothesis. Therefore, we can say that there is sufficient evidence to conclude that the relationship between X an Y could be linear, i.e. $H_0 : \rho > 0$, agreeging with the previous case.

Spearman

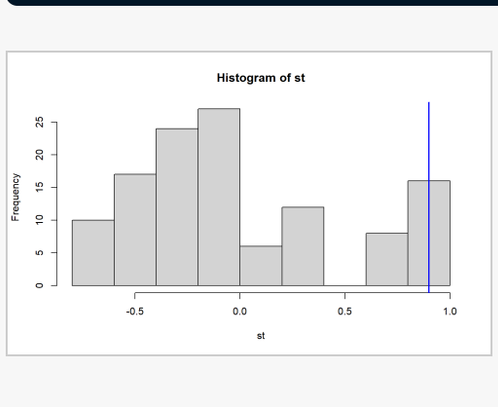
Finally, will now follow the same strategy but using Spearman's correlation coefficient instead of Pearson's.

```
x= c(39,29,60,40,32)
y=c(11,5,20,8,6)

sttrue= cor(x,y, method = "spearman")

n=length(y)
nr=fact(n) #number of rearrangements to be examined
st=numeric(nr)

cnt=0
d=permn(y)

for (i in 1:nr){
  st[i]<-cor(d[[i]],x)
  if (st[i] >=sttrue)
    cnt=cnt+1
}
```

```
print(paste("p-value= ",cnt/nr))
```

```
## [1] "p-value=  0.0666666666666667"
```

```
hist(st)
abline(v=sttrue,col="blue",lwd=2)
```



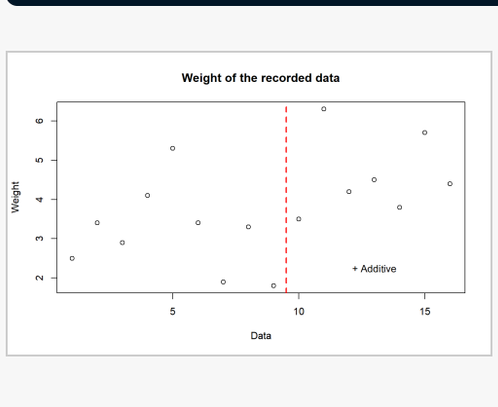
In this case the p-value $= 0.066$ which is slightly greater than \$ 0.05\$. Therefore, using the Spearman's statistic we could have concluded that H_0 is likely and that we cannot reject it, however, the p-value is quite low and the other statistics rejected the null hypothesis.

In conclusion, even thought the null hypothesis will be rejected using Spearman's statistic we can conclude due to the other permutations tests that X and Y are correlated.

TASK 2

We have to analyse the increments of weight recorded with the new additive. For that purpose, firstly we will introduce the data in R and look at it.

```
data_ex2 <- c(2.5, 3.4, 2.9, 4.1, 5.3, 3.4, 1.9, 3.3, 1.8)
data_additive_ex2 <- c(3.5, 6.3, 4.2, 4.5, 3.8, 5.7, 4.4)

plot(c(data_ex2, data_additive_ex2), xlab = "Data", ylab ="Weight", main = "Weigh
abline(v = 9.5, col="red", lwd = 2, lty = 2)
text(13, 2, "+ Additive", srt=0.2, pos=3)
```



Looking at this plot it seems that the weight had grow when using the new additive. Furthermore, we will do an statistical analysis to prove our feelings.

We will do a t-test with

$$H_0 : \mu_{no_a dditive} = \mu_{additive}$$

against

$$H_1 : \mu_{no_a dditive} < \mu_{additive}$$

However, for doing this test we need to check if the variances are equal, so we will first do the following test:

$$H_0 : \frac{\sigma_A^2}{\sigma_B^2} = 1$$

$$H_1 : \frac{\sigma_A^2}{\sigma_B^2} \neq 1$$

```
library(car)
```

```
## Loading required package: carData
```

```
total_data_ex2 = stack(list(g1=data_ex2, g2=data_additive_ex2))
leveneTest(values ~ ind, total_data_ex2)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.0437 0.8374
##       14
```

Notice that our p-value of the test is 0.8374 > 0.05, so we will accept the Null hypothesis and conclude that the variance are equal.

Now, we also have to check if the data follows a normal distribution, for that purpose we will do a *Shapiro-Wilk* test:

```
shapiro.test(data_ex2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_ex2
## W = 0.94362, p-value = 0.6205
```

```
shapiro.test(data_additive_ex2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_additive_ex2
## W = 0.90713, p-value = 0.3763
```

On the 2 cases we have that the p-value > 0.05 and then it implies that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality.

Finally, we can do the t-test:

```
t.test(data_ex2, data_additive_ex2, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  data_ex2 and data_additive_ex2
## t = -2.7468, df = 13.485, p-value = 0.008093
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.5179979
## sample estimates:
## mean of x mean of y
##  3.177778  4.628571
```

Due to the p-value < 0.05, we reject the null hypothesis and we have that $\mu_{no_a dditive} < \mu_{additive}$.

Now we can perform a permutation test with the T-statistic using the following R command *perm.t.test(x,...)* from the package *MKinfer*,

```
library(MKinfer)
```

```
## Warning: package 'MKinfer' was built under R version 4.1.2
```

```
perm.t.test(data_ex2, data_additive_ex2, alternative = "less")
```

```
##
##  Permutation Welch Two Sample t-test
##
## data:  data_ex2 and data_additive_ex2
## (Monte-Carlo) permutation p-value = 0.008601
## 95 percent (Monte-Carlo) permutation percentile confidence interval:
##        -Inf -0.4126984
##
## Results without permutation:
## t = -2.7468, df = 13.485, p-value = 0.008093
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.5179979
## sample estimates:
## mean of x mean of y
##  3.177778  4.628571
```

Notice that we also obtain a p-value < 0.05, so will reject the null hipotesys and accept that $\mu_{no_a dditive} < \mu_{additive}$.

Furthermore, we can do more permutation test. Due to having data of different size we will use the following idea for doing a permutation test: 1. Add two NA values to the additive data and sample the data. 2. Delete two element of the data, the ones that correspond to the NA introduced data. 3. Do the permutation test. 4. Compute an average of the p-values of the permutation tests.

```
iter <- 10
p_value_count <- 0

for (j in 1:iter){
  mod_data_additive_ex2 <- c(data_additive_ex2, NA, NA)
  mod_data_additive_ex2 <- sample(mod_data_additive_ex2)
  sttrue= cor(data_ex2, mod_data_additive_ex2, use="complete.obs")

  n=length(mod_data_additive_ex2)
  nr=fact(n) #number of rearrangements to be examined
  st=numeric(nr)
  st <- 0
  cnt=0
  d=permn(mod_data_additive_ex2)

  for (i in 1:nr){
    st[i]<-cor(d[[i]],data_ex2, use="complete.obs")
    if (st[i] >=sttrue)
      cnt=cnt+1
  }
  print(paste("p-value= ",cnt/nr))
  p_value_count <- p_value_count + cnt/nr
}
```

```
## [1] "p-value=  0.293656305114638"
## [1] "p-value=  0.410449735449735"
## [1] "p-value=  0.929557980599647"
## [1] "p-value=  0.596924603174603"
## [1] "p-value=  0.682991622574956"
## [1] "p-value=  0.101085758377425"
## [1] "p-value=  0.91074184303351"
## [1] "p-value=  0.67949184303351"
## [1] "p-value=  0.817708333333333"
## [1] "p-value=  0.913944003527337"
```

```
print(paste("p-value= ",p_value_count/iter))
```

```
## [1] "p-value=  0.633655202821869"
```

```
iter <- 10
p_value_count <- 0

for (j in 1:iter){
  mod_data_additive_ex2 <- c(data_additive_ex2, NA, NA)
  mod_data_additive_ex2 <- sample(mod_data_additive_ex2)
  sttrue= mean(mod_data_additive_ex2, na.rm = TRUE) - mean(data_ex2)

  n=length(mod_data_additive_ex2)
  nr=fact(n) #number of rearrangements to be examined
  st=numeric(nr)
  st <- 0
  cnt=0
  d=permn(mod_data_additive_ex2)

  for (i in 1:nr){
    st[i]<-mean(d[[i]], na.rm = TRUE) - mean(data_ex2)
    if (st[i] >=sttrue)
      cnt=cnt+1
  }
  print(paste("p-value= ",cnt/nr))
  p_value_count <- p_value_count + cnt/nr
}
```

```
## [1] "p-value=  1"
## [1] "p-value=  1"
## [1] "p-value=  1"
## [1] "p-value=  1"
## [1] "p-value=  1"
## [1] "p-value=  1"
## [1] "p-value=  1"
## [1] "p-value=  1"
## [1] "p-value=  1"
## [1] "p-value=  1"
```

```
print(paste("p-value= ",p_value_count/iter))
```

```
## [1] "p-value=  1"
```

As we can see, the average p-value for both test is greater than0.05, so we will reject the null hypothesis and conclude that in one case that are correlated and in the other that $\mu_{no_a dditive} < \mu_{additive}$ as we had concluded with the t-test and the permutation test done before.

Notice that similar test could be carried with different statistics such as Spearman or Pearson.

TASK 3

The first thing we will do is to create a Data frame with the following variables: "TEMP" (Temperature), "TMG" (Time mowing the grass) and "WC" (Water Consumption). Then, we will use the function "lm" in R to build a model that predicts the variance of "WC" as a function of the other two previously defined variables. So,

```
df<-data.frame(TEMP = c(75,83,85,85,92,97,99),
          WC = c(16,20,25,27,32,48,48),
          TMG=c(1.85,1.25,1.5,1.75,1.15,1.75,1.6))


model <- lm(WC ~ TEMP+TMG,data=df)
summary(model)
```

```
##
## Call:
## lm(formula = WC ~ TEMP + TMG, data = df)
##
## Residuals:
##       1       2       3       4       5       6       7
##  1.0441  0.4642 -0.6935 -1.8264  0.1061  1.0252 -0.1197
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -121.65500    6.54035 -18.601 4.92e-05 ***
## TEMP           1.51236    0.06077  24.886 1.55e-05 ***
## TMG           12.53168    1.93302   6.483  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.245 on 4 degrees of freedom
## Multiple R-squared:  0.9937, Adjusted R-squared:  0.9905
## F-statistic: 313.2 on 2 and 4 DF,  p-value: 4.027e-05
```

As we can see, both p-values for the variables *TEMP* and *TMG* have a values lower than 0.05. Therefore, we can conclude that both variables are significant in explaining the variation of *WC*. However, if we compare these two values we can see that in the case of *TEMP* the value is 188 times lower. This means that despite having two significant variables, one will have more weight than the other.

On the other hand, if we look at the coefficients of the regression we can see that the values of *TEMP* are multiplied by 1.512 while those of *TMG* are multiplied by 12.531. It may seem strange that the coefficient of the variable *TEMP* is lower than that of *TMG*. However, it should be noted that the mean of *TEMP* is approximately 56 times higher than that of *TMG*.

Once we have obtain the results from the model we will compare de p-values using the correlation test.

```
# CORRELATION TEST

x=df$TEMP
y=df$WC

sttrue= cor(x,y)

n=length(y)
nr=fact(n) #number of rearrangements to be examined
st=numeric(nr)

cnt=0
d=permn(y)

for (i in 1:nr){
  st[i]<-cor(d[[i]],x)
  if (st[i] >=sttrue)
    cnt=cnt+1
}
```

```
print(paste("p-value= ",cnt/nr))
```

```
## [1] "p-value=  0.000793650793650794"
```

```
library(combinat)
x=df$TMG
y=df$WC

sttrue= cor(x,y)

n=length(y)
nr=fact(n) #number of rearrangements to be examined
st=numeric(nr)

cnt=0
d=permn(y)

for (i in 1:nr){
  st[i]<-cor(d[[i]],x)
  if (st[i] >=sttrue)
    cnt=cnt+1
}
```

```
print(paste("p-value= ",cnt/nr))
```

```
## [1] "p-value=  0.41031746031746"
```

As we can observe in the case of the variable *TMG*, the p-value is greater than 0.05. Therefore, we cannot discard the null hypothesis (there is no correlation between *TMG* and *WC*). In other words, we cannot say that there is a correlation between these two variables.

On the other hand, the variable "TEMP" does present a p-value lower than 0.05. Therefore, we can discard the null hypothesis and we can state that there is a correlation between the two variables.

Notice that this results agree with the ones obtained with the regression model. Indeed, we have seen in the regression model that the variable *TEMP* is more significant than *TMG*, and using the correlation tests we also notice that *TEMP* is correlated with *WC* which support the results from the regression model.